

Practical Limitations in Making Decisions Regarding the Distribution of Applicant Personality Test Scores Based on Incumbent Data

Jennifer P. Bott · Matthew S. O'Connell ·
Mano Ramakrishnan · Dennis Doverspike

Published online: 11 July 2007
© Springer Science+Business, LLC 2007

Abstract For practitioners, the possibility of faking on personality tests has potential implications that are much broader than those captured by current theoretical debates over criterion-related validity, factor structure, or psychological processes. One unexplored potential impact of response distortion involves the pass rates associated with applying cutoff scores developed using a concurrent validation design to applicant samples. This practitioner-oriented paper compared applicant and incumbent scores on three personality dimensions and uncovered significant standardized group differences. These differences greatly influenced pass rates for three different selection models, which impacted expected utility of the selection system. Potential solutions for practitioners are provided, along with recommendations for future research in this area.

Keywords Response distortion · Faking · Personality tests · Utility · Cutoffs · Scores · Selection

An earlier version of this paper was presented at the annual meeting of the Society for Industrial and Organizational Psychologists in Orlando, Florida.

J. P. Bott (✉)
Department of Marketing and Management, Miller College
of Business, Ball State University, Muncie, IN 47306, USA
e-mail: jpbott@bsu.edu

M. S. O'Connell
Select International, Inc., Pittsburgh, PA, USA

M. Ramakrishnan
George Washington University, Washington, DC, USA

D. Doverspike
Department of Psychology, University of Akron, Akron,
OH, USA

Introduction

Concurrent validation designs, in which both predictor and criterion data are collected from job incumbents, are often used by human resource practitioners for the purpose of establishing operational selection systems (e.g., Smith et al. 2001; Stewart and Carson 1995). While this strategy is typically more efficient and cost-effective, experts in the field recognize a number of problems with concurrent designs, including restriction of range and motivational differences (Barrett et al. 1981; Cascio 1998; Guion and Cranny 1982). The negative outcomes associated with motivated responding may become even more salient when personality tests are the object of validation, as these measures are more transparent and susceptible to response distortion than cognitive ability measures (e.g., Alliger et al. 1996).

Empirical research has consistently demonstrated that job applicants (both in the laboratory using instructional sets and in field settings) distort responses on personality measures to appear more qualified for a position (e.g., Donovan et al. 2003; Ellingson et al. 1999). Mounting empirical evidence has demonstrated that applicants increase their scores on non-cognitive tests by approximately a half standard deviation (Rosse et al. 1998; Viswesvaran and Ones 1999). Empirical and theoretical work in this area has primarily examined the resultant criterion-related validity of the selection system (e.g., Douglas et al. 1996; Hough et al. 1990; Mueller-Hanson et al. 2003; Ones et al. 1996). Many studies have uncovered large differences between responses from applicants and incumbents; few studies have examined the practical implications of these differences. No study has yet to examine how response distortion influences the application of cutoff scores established using a concurrent validation design onto an applicant sample.

The present study investigates the impact of response distortion on the application of cutoff scores developed using a concurrent design. Avoiding the ongoing debate over validity, we focus on two practical issues: (1) What are the effects of using incumbent data from a concurrent study to set cutoff scores on personality tests for an applicant sample? and (2) What might account for elevated applicant scores: Social desirability, cognitive ability, or both?

Personality Assessment and Response Distortion

Non-cognitive measures, which include personality assessments, are frequently used in selection systems for two primary reasons: (1) Using job-relevant personality constructs, like conscientiousness, can add incremental predictive validity above and beyond cognitive ability (Barrick and Mount 1991; Ones et al. 1993) and (2) Personality assessments often result in little or no adverse impact (Ones et al. 1993; Cunningham et al. 1994). However, from a practitioner perspective, the main disadvantage of personality assessment is its susceptibility to response distortion (e.g., Douglas et al. 1996; Stark et al. 2001).

As briefly presented above, response distortion, or the impression management dimension of socially-desirable responding, refers to intentional distortion of responses that can be situationally-induced, as in a job setting (e.g., Ellingson et al. 1999; Paulhus 1984). Research has shown that individuals, when instructed to do so in a laboratory setting, can increase their responses by approximately a half standard deviation (e.g., Viswesvaran and Ones 1999). This inflation has been replicated in field settings, where individuals are actual job applicants (e.g., Rosse et al. 1998).

Much of the current debate regarding response distortion has attempted to determine if these elevated responses impact organizational outcomes, including the criterion-related validity of the selection system (e.g., Christiansen et al. 1994; Douglas et al. 1996; Mueller-Hanson et al. 2003; Ones et al. 1996). Less research has examined the influence of response distortion on measurement properties, including factor structures (e.g., Schmit and Ryan 1993; Van Iddekinge et al. 2001) and quality of selection decisions (e.g., Christiansen et al. 1994). Studies have primarily focused on establishing the boundary conditions for response distortion. One boundary condition that has yet to be investigated is the impact on cutoff scores established using job incumbents in a concurrent validation approach.

It is our contention that applicants have more of a reason to distort their responses than job incumbents (i.e., attempting to secure a desired position within an organization). As a result, we believe the average score of

applicants on personality tests would be higher than the average score of incumbents. We believe this will influence the number and percentages of individuals passing tests with minimum cutoffs established using job incumbents. Before presenting the purpose of the current study, we briefly review the literature related to cutoff scores.

Cutoff Scores

A cutoff score corresponds to a point or score on a test below which a person is considered to have failed the test or selection device (Cascio et al. 1988; Hoffman and Thornton 1997). This is different than a critical score which is “a specified point in a distribution of scores at or above which candidates are considered successful...” (SIOP 1987, p. 37). Cutoff scores are often used within a non-compensatory system, such that individuals must pass each step in a multiple cutoff selection system. Although subjective methods can be used to determine cut scores (for a discussion, see Cascio et al. 1988), practitioners frequently make use of local norms, which refer to normative information (e.g., means, standard deviations) for the applicant or incumbent population of interest (Kehoe and Olson 2005; Crocker and Algina 1986).

If incumbents are used to establish cut scores, as is commonly done in a concurrent validation approach (Kehoe and Olson 2005), and if applicant responses are distorted upwards, then data from incumbents may seriously underestimate the cutoff scores required for applicants. This potential unexpected increase in pass rates could influence the overall efficiency of the selection system, as well as its utility.

Explaining Response Distortion: Social Desirability or Cognitive Ability

As is apparent in the response distortion literature, individuals can inflate responses both in laboratory settings with different instructional sets and in field settings with different motivation. However, all individuals do not distort to the same degree. One possible rationale for this inconsistency is intelligence, such that those higher in cognitive ability are better able to identify effective strategies for response inflation. As such, theoretical models (e.g., McFarland and Ryan 2000; Snell et al. 1999) have proposed that ability to fake is an important determinant of response distortion. However, few studies have examined cognitive ability in an effort to predict personality scores and, by extension, response distortion. The extant literature offers few guidelines, as only a handful of studies have examined cognitive ability's relation to personality scores. For example, in a laboratory study which varied instructional sets and examined integrity test performance, Alliger

et al. (1996) found that correlations between intelligence scores (as measured by the Shipley Institute of Living Scale) and integrity test scores were highest in the conditions which received fake good and coaching instructions. The authors concluded that intelligent individuals were more likely to respond to the instructional sets in an effective manner. This finding contradicts the results of Werner et al. (1989), in which no relationship was uncovered between two measures of intelligence (i.e., education level, scores on Scale B of the 16PF) and integrity test scores.

A second goal of this study, then, was to understand whether cognitive ability or social desirability scores better predict personality scores in order to identify the likely determinants of response distortion. The results of this field study will shed light on these apparent inconsistencies and offer empirical evidence of the importance of cognitive ability in the prediction of personality scores.

Current Project

The data described here were collected as part of an applied project involving the introduction of a selection system, which included a personality test. The project involved collecting data from an incumbent sample in order to estimate the effects of personality test use. Applicant data were collected at a later date, allowing for a comparison of the results from the concurrent study to the results for the applicants.

With this applied sample, we were interested in investigating the practical utility of setting cut scores for an applicant sample based on incumbent data. In order to examine this adequately, we first determined if differences between datasets existed, as well as the potential explanation for those differences. Specifically, we were interested in ruling out rival explanations for the personality score differences, including cognitive ability. After establishing the likely cause of these differences, we turned to the focal analysis: Pass rates related to incumbent-based cutoff scores using three different models.

Method

Participants

Incumbent Dataset

The participants were 303 current employees selected using a stratified, random sampling strategy from two primary business units of a large manufacturer in the U.S. All participants were hourly workers who had been with the

company for at least 6 months. Although there were multiple job titles covered in the sample, including packer, material handler, and operator, all of the positions were hourly production positions that shared the same job classification. These respondents had been employed for an average of 13.27 years ($SD = 10.25$). While an average of 13 years in hourly positions may seem high, it is important to take into account that the current organization was seen as an employer of choice that typically pays at the 85th or higher percentile in the communities where the plants were located. The annual voluntary turnover rate at this organization was less than 5%. Incumbents were paid their normal hourly wage for participation in the study, which took them approximately 2 h.

Depending on the time they entered the organization, the employees in the incumbent dataset were selected using a combination of an application screen, interview focusing on past work experiences, and a background check for criminal activity; more recently a cognitive ability test with a very low cutoff was added. Thus, there was relatively little reason to believe there was much restriction of range in the incumbent dataset, and this assumption was supported by a comparison of the cognitive ability scores of the incumbent and applicant datasets (see Table 2).

Applicant Dataset

Data came from applicants to the same organization, gathered between 2000 and 2004, applying for the same hourly manufacturing positions covered in the incumbent dataset. In total, there were 5,629 individuals in the dataset. Average age for this sample was 36 years ($SD = 10.09$). A majority of the sample were white (67.5%), followed by African-American (21.8%). Most respondents were men (68%). Prior to completing the testing process, these applicants were screened based on a minimal set of qualifications established by the organization, including willingness to work shifts and acceptance of the hourly wage. All applicants completed the testing process during on-site proctored administrations. Testing conditions were similar to those used with the incumbent group.

Measures

Cognitive Ability

A 35-item computer-administered cognitive ability measure was used. Items were written to reflect three broad domains, including analytical, numerical, and applied reasoning. The sum across the 35 items was calculated and then converted into a stanine-type score on a 1–10 scale. O'Connell and Kato (2001) reported an internal consistency reliability of .81 based on a sample of 3,311 individuals. Meta-analysis

by O'Connell (2000), based on 7 manufacturing samples and 718 individuals, reported a corrected correlation of .43 with supervisor's performance ratings. This meta-analytic coefficient was corrected for both range restriction and unreliability of measures.

Personality

The tests used in the study were all part of a computer-based assessment system designed for use in a broad range of manufacturing occupations. Descriptions of parts of this computer-based assessment system, the Select Assessment for Manufacturing,TM have appeared elsewhere in the literature (Hattrup et al. 2005; O'Connell et al. 2001; O'Connell et al. 2002). All participants completed the same assessment battery in proctored environments. The test battery consisted of a cognitive test, several personality tests, and other simulations. The results reported here are for the three personality tests. All meta-analytic criterion-related validity coefficients presented below were corrected for range restriction and unreliability of measurement.

All personality items were single statements to which the respondent used a sliding pointer to indicate agreement or disagreement on a 0 to 100 scale. The responses were then converted to continuous variables with a range from 1 to 5 (Strongly Disagree to Strongly Agree). A similar graphic slider scale was described in Cook et al. (2001). Each of the individual personality scales is described in more detail below.

Agreeableness

Agreeableness was measured using a 15-item scale with an internal consistency reliability reported by O'Connell and Kato (2001) of .72. Coefficient alpha for the incumbent sample was .76 and .73 for the applicant sample. Meta-analysis by O'Connell (2000) based on seven manufacturing samples and 718 individuals reported a correlation of .28 with supervisor's performance ratings. Normative data was available for both applicant and incumbent respondents, as well as for race. Example items include, "There are a lot of people I don't get along with," and "I usually make friends easily."

Conscientiousness

A 23-item scale described by Hattrup et al. (1998), who reported an internal consistency of .70, was used to measure conscientiousness. Internal consistency for the incumbent dataset was .82 and .88 for the applicant dataset. The scale was correlated .23 and -.24 with organizational citizenship behaviors and absenteeism, respectively, among a sample of retail sales workers. In addition, the meta-analysis

described above (O'Connell 2000) found a corrected correlation of .44 for this scale with supervisor ratings of performance. Example items include, "I hate wasting time," and "I enjoy tackling difficult tasks."

Emotional Stability

A 10-item scale was used to measure negative affectivity; however, prior to analysis, all items were reverse coded. For ease of interpretation, this dimension was called emotional stability, with higher scores indicating higher levels of emotional stability. Prior research found an internal consistency reliability of .71 (O'Connell and Kato 2001). This scale had a corrected correlation of .33 with supervisor ratings (O'Connell and Smith, 1999). Example items include, "When things go wrong, I blame myself," and "I always expect the worst to happen." Coefficient alphas for the incumbent and applicant dataset was .67 and .74 respectively.

Social Desirability

A 10-item short version of the Crowne–Marlowe (1960) social desirability scale, developed by Strahan and Gerbasi (1972), was used. Internal consistency for this measure was .60 in the incumbent and .68 in the applicant sample. Research by Fischer and Fick (1993) reported a correlation of .96 with the 33-item version of the Crowne–Marlow scale and an internal consistency reliability of .88.

Results and Practical Implications

We performed several analyses in order to evaluate our questions of interest and eliminate possible rival explanations for our findings. We began by examining descriptive information about the variables of interest in both datasets, followed by a presentation of standardized mean differences across datasets. This serves to establish the level of response distortion on personality tests between the applicant and incumbent groups. We next examined the two predictors of personality differences, namely social desirability and cognitive ability. We conclude with a discussion of a practical issue related to these personality differences, setting cut scores for the applicant dataset using incumbent data.

Descriptive Information

Table 1 presents the means, standard deviations, correlations, and internal consistencies for the focal variables, categorized by dataset type. As can be seen for both datasets, the three personality variables were moderately to

Table 1 Means, standard deviations, reliabilities and intercorrelations, organized by dataset

Incumbent dataset ^a								
Variables	Mean	SD	1	2	3	4	5	6
1. Agreeableness	3.61	.41	(.75)					
2. Conscientiousness	4.04	.37	.35**	(.83)				
3. Emotional stability	3.63	.45	.39**	.44**	(.67)			
4. Tenure ^b	159.19	122.97	-.09	-.20**	-.02	–		
5. Cognitive ability	5.06	1.33	-.03	.06	-.01	-.27**	–	
6. Social desirability	3.46	.42	.41**	.44**	.32**	-.06	-.01	(.60)
Applicant dataset ^c								
Variables	Mean	SD	1	2	3	4	5	
1. Agreeableness	3.98	.35	(.72)					
2. Conscientiousness	4.42	.36	.45**	(.89)				
3. Emotional stability	3.91	.42	.37**	.62**	(.70)			
4. Cognitive ability	4.84	1.39	.08**	.13**	.18**	–		
5. Social desirability	3.92	.43	.45**	.59**	.48**	.09**	(.66)	

^a N = 303 for all analyses, except cognitive ability, where N = 302

^b Calculated in months

^c N = 5,629 for all analyses

* $p < .05$ ** $p < .01$

highly intercorrelated (r ranges from .35 to .44 for the incumbent dataset; .37 to .62 for the applicant dataset). This is consistent with prior research that established overlap within measures of the Big Five in field samples (e.g., Costa and McCrae 1992). For the incumbent sample, data were available on tenure. Tenure was negatively correlated with conscientiousness ($r = -.20$, $p < .01$) and negatively related to cognitive ability ($r = -.27$, $p < .01$).

Applicant-incumbent Differences

After these preliminary analyses, we examined whether group differences existed based on the dataset under investigation. Specifically, we examined whether there were group mean differences on the cognitive ability and personality variables between the applicant and incumbent datasets. If there was no response distortion on the personality variables, one might expect incumbents to score higher than applicants, since they had been somewhat pre-selected, albeit using interviews as opposed to a personality test. Arguably, interviews may have been used as an indirect measure of personality, as well as previous job experience and biodata items. Therefore, it would make sense that those with personalities that “fit” the organization would have been hired, thereby leading to the possibility of higher incumbent scores. On the other hand, if applicants were distorting their responses in an attempt to fake good, one would expect the responses of applicants to be higher than incumbents. This may take place because

incumbents do not feel there is any value to distorting scores because they already have the job, whereas applicants may distort responses in order to appear more fitting for the position.

Table 2 presents the means, standard deviations, and mean standardized incumbent—applicant differences (Cohen 1988). An inspection of Table 2 reveals no significant mean difference between the applicant and incumbent datasets on cognitive ability ($d = .16$). Thus, there appears to have been relatively little restriction of range of cognitive ability scores in the incumbent dataset. In order to more fully examine the issue of range restriction, we performed independent sample t-tests for each of our focal variables in order to determine if the variances between the datasets were significantly different. The variances for the applicant and incumbent dataset were only significantly different for one variable, agreeableness (Levene’s test for homogeneity of variance; $F = 5.93$, $p < .05$). The other two personality tests, as well as cognitive ability, did not yield any variance differences between datasets. Therefore, we feel confident that range restriction is not unduly influencing our results. This lack of range restriction is similar to that which was found by Ones and Viswesvaran (2003) in their comparison of variability in applicant pools to national norms.

There were significant dataset differences, however, between the incumbent and applicant samples for the personality variables. As illustrated in Table 2, the means for applicants were statistically and practically higher than the

Table 2 Means, standard deviations, and standardized differences organized by category (incumbent vs. applicant)

Variable	Incumbent mean ^a	Incumbent SD	Applicant mean ^b	Applicant SD	Mean standardized difference (<i>d</i>)
Cognitive ability ^c	5.06	1.33	4.84	1.39	.16
Conscientiousness	4.04	.37	4.38	.36	1.05*
Agreeableness	3.61	.41	3.97	.36	1.05*
Emotional stability	3.62	.45	3.87	.43	.66*

^a Incumbent dataset *N* = 303^b Applicant dataset *N* = 5,629^c For cognitive ability calculations in the incumbent dataset, *N* = 302**p* < .05

means for incumbents for the three personality scales. An examination of Table 2 reveals the standardized difference between applicants and incumbents was 1.05 for conscientiousness, 1.05 for agreeableness, and .66 for emotional stability. In two of the three cases, the standardized difference between datasets was over a full standard deviation. These large standardized differences were greater than significant results reported on all five personality dimensions (and facets) by Rosse and colleagues (1998). These results, then, support the notion that applicants inflated their scores, for in all three cases, their scores were significantly higher than the incumbents. Assuming that no differences inherently exist between applicants and incumbents in true personality or cognitive ability, the consistently higher scores for the applicant group on conscientiousness, agreeableness, and emotional stability likely represent motivation to appear more qualified for the position.

Predicting Personality Using Social Desirability

In addition to the establishment of significant differences between datasets in terms of personality scores, we also investigated what individual differences (i.e., cognitive ability and social desirability) contributed to the prediction of personality scores using regression. Table 3 presents the results of this regression predicting personality composite scores, a unit-weighted average of the three personality dimensions, from social desirability, dataset, and their interaction. As can be seen in the table, dataset (applicant versus incumbent, coded 0 and 1, respectively) accounted for significant variance in both step one ($\beta = -.26$, $p < .001$) and step two ($\beta = -.11$, $p < .001$). Social desirability, entered at step two, reduced the effect of dataset and explained an additional 29% of the variance in personality composite scores. When the interaction between social desirability and dataset was entered at step three, the effect of dataset became non-significant ($\beta = .10$, *ns*). At the final step, the main effect for social desirability score

Table 3 Predicting personality composite scores using social desirability and dataset

Model	Variables	β	<i>t</i>	ΔR^2
1	Dataset ^a	-.26	-21.10***	.07***
2	Dataset	-.11	-10.49***	
	Social desirability	.62	61.06***	.36***
3	Dataset	.10	1.18	
	Social desirability	.62	59.79***	
	Social desirability by dataset	-.20	-2.50*	.00

N = 6,012^a Dataset was dummy-coded, such that Applicant = 0, Incumbent = 1**p* < .05, ***p* < .01, ****p* < .001

remained significant ($\beta = .62$, $p < .001$) and the interaction between dataset and social desirability became significant ($\beta = -.20$, $p < .05$).

These results indicated that social desirability was the largest contributor to unit-weighted personality composite. The interaction, while significant, did not add any additional variance beyond that of dataset and social desirability, calling into question the practical significance of this interaction. One method for determining the practical significance of a statistically significant finding is to examine measures of association (Svyantek and Ekeberg 2001), one of which is the percent of variance accounted for by the additional predictor (in this case, the interaction). As this additional variance was zero, we believe the interaction lacks practical significance. This statistically significant interaction was likely the result of the large sample size (*N* = 6,012). Overall, these results indicated that response distortion is driving dataset differences in personality composite scores.

Effect of Cutoffs

Practitioners frequently set cut scores by looking at the distribution of scores from local norms established based

upon the incumbent group (c.f. Cascio et al. 1991; Kehoe and Olson 2005). In some cases, they may also use criterion data and set the cutoff in some manner so as to distinguish between high and low performers. As Cascio et al. (1988) stated, “Since it is unlikely that an accurate base rate of job performance can be determined for the applicant population, it will be necessary to extrapolate either from performance data on all persons hired or from the present workforce (p. 17)”. In order to investigate the effects of setting a cutoff based upon incumbent personality data, we established three cutoff scores. We do not mean to imply that any of these methods represent the optimal methodology; our purpose here is to evaluate the impact of setting scores based on the incumbent group on the selection of applicants, which is a common practice (e.g., Pulakos et al. 2002).

The three cut scores were set rationally. The first cut score was set at the *mean* of the incumbent group (4.04 for conscientiousness, 3.61 for agreeableness, and 3.63 for emotional stability). The second cut score was set at *one standard deviation below the mean* of the incumbent group (3.67 for conscientiousness, 3.20 for agreeableness, and 3.18 for emotional stability). The third cut score, *one half standard deviation*, was set to fall between the mean and one standard deviation below the mean (3.85 for conscientiousness, 3.40 for agreeableness, and 3.40 for emotional stability).

Once the cut scores were established, they were applied to the personality test results from both the incumbent and applicant datasets. We calculated the results using a multiple cutoff scenario, (i.e., the test taker had to pass all three personality tests), as well as two compensatory models. For the compensatory model with cognitive ability, a unit-weighting strategy was employed. Therefore, we placed the cognitive ability and personality test scores on the same metric and summed them. For the personality only compensatory system, the three personality scores were simply summed. To pass in the compensatory system, the test taker had to score above the overall mean cutoff for all tests. The pass rates for each personality test, the multiple cutoff, and compensatory scenarios are presented in Table 4. Figure 1a–c also graphically represent the difference in pass rates between applicants and incumbents in the multiple hurdle and compensatory systems.

Inspection of Table 4 reveals that the results were quite different in the applicant and incumbent group. For example, one would expect that setting a cutoff at the mean would screen out approximately 50% of the individuals taking any single test (as the mean acts as a fulcrum for the distribution, assuming a normal distribution). Across the personality tests and in the compensatory systems, this was essentially what we found when examining the incumbent dataset. In the multiple cutoff situation, approximately 22%

Table 4 Pass rates for cutoff scores by dataset for personality variables individually and within a multiple cutoff and compensatory system

Source	Mean cutoff (%)	SD cutoff (%)	Half SD cutoff (%)
<i>Conscientiousness</i>			
Incumbent ^a	51.0	86.5	71.1
Applicant ^b	84.0	97.5	94.1
<i>Agreeableness</i>			
Incumbent	51.0	84.1	71.6
Applicant	85.9	98.3	95.3
<i>Emotional stability</i>			
Incumbent	49.5	86.2	71.1
Applicant	74.0	96.0	88.4
<i>Multiple cutoff^c</i>			
Incumbent	22.4	68.5	46.1
Applicant	62.2	93.4	83.1
<i>Compensatory—cognitive ability and personality^d</i>			
Incumbent	52.3	84.9	71.0
Applicant	63.2	100	77.4
<i>Compensatory—personality only</i>			
Incumbent	52.3	83.9	70.3
Applicant	86.8	98.3	95.1

^a N = 303

^b N = 5,629

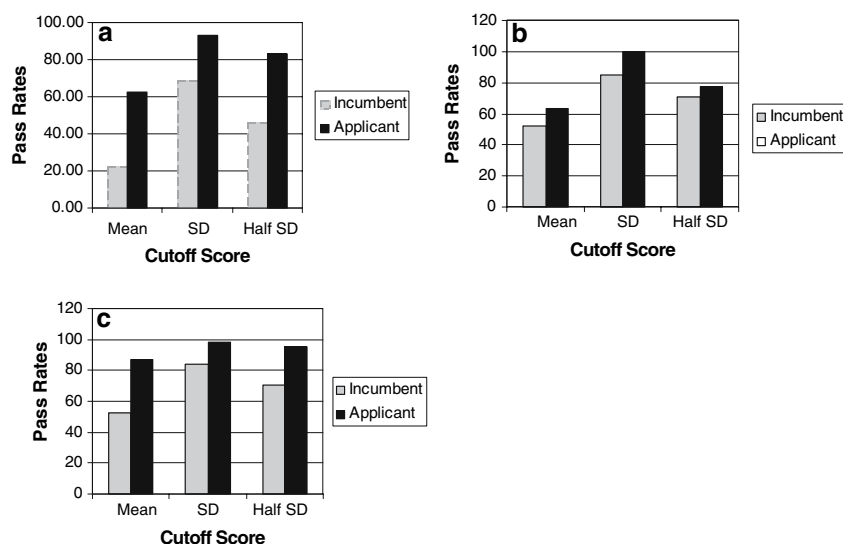
^c To be selected, respondents must meet a minimum score for agreeableness, conscientiousness, and emotional stability

^d To be selected, respondents must meet a minimum score for the unit-weighted average of agreeableness, conscientiousness, and emotional stability, and cognitive ability

of incumbent dataset passed. However, the results were drastically different when the mean cutoff was applied to the applicant group. For both conscientiousness and agreeableness, over 80% of the applicants would have passed. Over 70% would have passed using the mean for emotional stability. This greatly exceeds our expectations of a pass rate of 50%. For the applicant group, 62% passed the multiple cutoff system, which is dramatically different than the 22% rate for incumbent group (see Fig. 1a). Pass rates were also higher for applicants in both the compensatory systems, which can be seen in Fig. 1b and c.

Similar results were obtained for both the one and one half standard deviation below the mean cutoff scores. For the one standard deviation below the mean cutoff, the passing rates for the incumbent group were 86.5% for conscientiousness, 84.1% for agreeableness, 86.2% for emotional stability, 68.5% for the multiple cutoff, and approximately 85% for both compensatory systems. Again, the pass rates were much higher for the applicant group. For the applicant group, using the one standard deviation below the mean cut score, the pass rates were 97.5% for

Fig. 1 (a) Graphical representation of pass rates for applicants and incumbents for multiple hurdle system. (b) Graphical representation of pass rates for applicants and incumbents for compensatory system (cognitive ability and personality). (c) Graphical representation of pass rates for applicants and incumbents for compensatory system (personality only)



conscientiousness, 98.3% for agreeableness, 96% for emotional stability, 93.4% for the multiple cutoff, 100% for the compensatory system with cognitive ability, and 98.3% with the personality only composite (see Fig. 1a–c).

For the one half a standard deviation cutoff score, the pass rates for the incumbent group were 71.1% for conscientiousness, 71.6% for agreeableness, 71.1% for emotional stability, 46.1% for the multiple cut score, 71% for the compensatory model with cognitive ability, and 70.3% with the personality compensatory system. As demonstrated previously, the pass rates were much higher for the applicant group. For the applicant group using the empirical cutoff, the passing rates were 94.1% for conscientiousness, 95.3% for agreeableness, 88.4% for emotional stability, 83.1% for the multiple cut score, 77.4% for the compensatory score with cognitive ability, and 95.1% for the personality compensatory composite (see Fig. 1a–c).

As these results demonstrate, there is a fairly drastic shift in pass rates when moving from the incumbent to the applicant dataset. The reliance upon the incumbent sample would lead to far more applicants passing the test than would be expected based upon the incumbent sample. This difference in pass rates would have major implications in terms of the costs, logistics, and time required in further screening of the passing applicants and could have long-term costs in terms of performance and turnover.

A related consideration is the operational utility of these personality tests. While not intending to present a comprehensive utility analysis, an illustrative example is instructive with respect to the impact that higher-than-expected pass rates on anticipated cost of a selection system. For instance, many practitioners use these personality measures to screen out individuals in a multiple cutoff process. If one assumes that the next step in the system would be an interview, one who based expectations on the

incumbent mean of the compensatory system would anticipate 2,944 individuals passing. However, pass rates from the applicant compensatory system with cognitive ability would yield 3,557 individuals passing the test. Using incumbent data in this scenario would result in 613 applicants more than expected (a 21% increase) passing the screening. With respect to this example, assuming an administrative cost of approximately \$100 per interview, this would cost the organization \$61,300 in additional selection system expenses.

Similarly, the pass rates and expected cost could be compared in a non-compensatory system (multiple cutoff). Based on the incumbent data, one would assume that 1,261 individuals would pass using the mean for each test as a cutoff. However, the pass rate is much higher than anticipated, resulting in 3,501 individuals moving onto the interview stage. Not only are the administrative costs of processing this many more applicants significant, the value of using a selection tool in which 80–90% of the applicants pass is extremely questionable. These two models, compensatory and non-compensatory, both illustrate a further problem associated with establishing cut scores using an incumbent sample, namely cost and utility.

Discussion

The purpose of the current study was to evaluate the effects of differences in response distortion between incumbents and applicants on the decisions practitioners make regarding test use. In particular, practitioners often use data from the results of a concurrent study, in which incumbents are administered the test, to make predictions concerning the operational use of the test with applicants for jobs. The types of decisions explored in this study primarily involved

the use of incumbent responses as a normative source for setting cutoffs.

The results clearly indicated that there were large differences in the mean responses and also the pass rates between incumbent and applicant groups. The standardized mean difference between the incumbent and applicant group were 1.05 for conscientiousness, 1.05 for agreeableness, and .66 for emotional stability. As a result of regression analyses, social desirability and, to a lesser extent, cognitive ability, predicted personality scores.

These differences are significant from an applied or practical standpoint in terms of their impact on expected pass rates. When data from the incumbent group was used to set cutoffs, the pass rates in the applicant group were much higher than would be expected. For example, using the mean cutoffs and a multiple cutoff approach resulted in 22.4% of the incumbent group passing. When the same data was applied to the applicant group, the passing rate was 62%.

These differences were found to cause serious problems in effectively utilizing results from an incumbent sample to estimate pass rates in an applicant sample. The results indicated the true pass rate in the applicant sample was more than twice that of the incumbent sample. However, as there were no significant differences for cognitive ability between the applicant and incumbent groups, the most likely explanation for the differences in the means on the personality tests was response distortion.

It should be noted that the results described above do not appear to be the result of restriction of range, an argument that would primarily involve the incumbent sample. Incumbents may exhibit only certain characteristics associated with successful performance at the organization and, over time, the homogeneity of incumbent characteristics may increase (c.f., the attraction-selection-attrition framework; Schneider 1987). However, we point to several indicators that range restriction in the incumbent sample is not the cause of these focal results. First, the incumbents had been selected based primarily upon an interview. While personality may be assessed informally during the interview process, the incumbents were not subject to written personality testing during the selection process. Therefore, the selection process used by the organization did not guarantee that only the top of the distribution of the three personality traits (conscientiousness, agreeableness, and emotional stability) were represented by incumbents. Second, if range restriction in the incumbent sample were present, one would expect the standard deviations for each of the three personality variables to be quite narrow; this was not the case (see Table 1, top half), as applicants appeared to exhibit similar responses (see Table 1, bottom half). Third, if the incumbents had been screened or selected based explicitly upon personality, then we would have expected their scores to be relatively elevated com-

pared to the applicants. Again, an inspection of Table 1 revealed this was not the case, as applicants had higher mean values for all three personality scales.

Potential Solutions for Practitioners

This paper highlights the practical problems that response distortion can have on setting cutoff scores and making selection decisions. Although there may not be a single remedy to this problem, there are some potential solutions that should be evaluated. First, some researchers have attempted to correct personality scores based on socially-desirable responding (e.g. Christiansen et al. 1994). Such an approach however does not typically result in improved criterion validity (Christiansen et al. 1994) and does not produce corrected scores that approximates honest scores (Ellingson et al. 1999). One reason why correcting for social-desirable responding does not improve criterion validity is that it may be related to important individual differences, such as conscientiousness and emotional stability (Ones et al. 1996).

A second potential solution is to provide warnings not to fake or, conversely, to be honest. In particular, warnings should indicate that faking can be identified and stress the negative consequences of faking (Dwight and Donovan 2003). Many theoretical models of faking (e.g., McFarland and Ryan 2000; Snell et al. 1999) indicated the importance of situational influences on response distortion. Future research should examine the manipulation of features of the situation as a potential remedy to response distortion.

Third, as suggested by Mueller-Hanson et al. (2003), personality scores could be used to “select out”, rather than “select in” applicants. Mueller-Hanson et al. (2003) demonstrated that using a “select in strategy”, in which those with high trait scores are expected to demonstrate high levels of job performance, will be more susceptible to response distortion, as this relationship between self-reported personality and performance does not hold for those distorting responses. However, using a “select out strategy,” in which low scorers are removed from the applicant pool and average to high scorers (and, by extension, possibly those who are distorting responses) are retained for further testing in which response distortion may not play such a large factor (e.g., situational judgment tests, cognitive ability tests, assessment centers). Arguably, while we cannot be certain if high scorers have high true scores or merely high response distortion, we can be relatively sure that individuals with low reported scores should be selected out, as low scores on these personality traits are not associated with high performance on the job (e.g., Barrick and Mount 1991; Saldago 1997).

Additionally, researchers should explore different types of personality tests that may be more resistant to faking.

For example, ipsative tests which present two socially desirable options (but only one which is job-relevant) can be presented. Unlike traditionally scales which use normative scores, ipsative scores for each person remains constant across the scale (e.g., Bartram 1996). Meaning, then, in an ipsative scale is determined by relative intra-individual comparison. As a result, some argue strongly against the use of ipsative measures for inter-individual comparisons (e.g., Hicks 1970; Johnson et al. 1988) as a result of interpretation and reliability issues, while others have found limited success with ipsative or partially-ipsative measures as a means to reduce the effect of faking at the group, rather than individual, level (e.g., Heggstad et al. 2006). Further research establishing the validity of the use of ipsative (or partially-ipsative) measures at the individual level (the level at which hiring decisions are made) is necessary before this becomes a viable option to limit the effects of response distortion.

Alternatively, personality tests can be presented in a situational judgment format, rather than a Likert self-report format. Some recent work has examined the relationship between implicit trait policies, personality, and situational judgment test response options, such that the endorsement of certain options reflects underlying personality traits (Motowidlo et al. 2006). Weekly et al. (2004) found that applicants scored lower on situational judgment test items than personality items (agreeableness, extraversion, conscientiousness), while retaining measurement equivalence and criterion-related validity across the applicant and incumbent samples. This research may provide promising avenues for assessing personality in more indirect way, thereby limiting the possibility and effect of response distortion.

Lastly, another possible solution would be to simply correct for the large applicant-incumbent difference in scores. However, there is little evidence to suggest that such differences are consistent across different samples and situations. For example, the standardized mean difference for agreeableness was 1.05 in our study. In contrast, Ployhart et al. (2003) reported a standardized mean difference for agreeableness of only .62. Rosse et al. (1998) found an average standardized mean difference of 1.09 between applicant and incumbent samples across measures of the Big Five. Additionally, Ployhart et al. (2003) suggested that such differences could vary according to test format. Future studies should include meta-analytic investigation of applicant-incumbent differences in order to better address this issue of the viability of corrections.

Future Research and Conclusions

Clearly, more research should be conducted to explore different options for addressing the substantial inflation of personality scores in applicant samples. Failure to recognize

issues related to response distortion raises serious logistical, cost, and resource issues in effectively utilizing these types of tests as part of a selection process. Future research should investigate appropriate methods for correcting, or at least addressing, the substantial inflation of personality scores in applicant samples.

In addition, failure to address this problem may raise serious legal concerns with regard to how cutoffs are set using an incumbent sample and applied to an applicant sample. In the current example, to realize a 50% pass rate in the applicant sample the corresponding cutoff would result in a pass rate of only 18% for conscientiousness in the incumbent sample, 20% for agreeableness, and 30.6% for emotional stability. These extremely low pass rates in the incumbent sample would be difficult to justify unless one were to admit that people can significantly inflate their scores on the personality measures, and therein call into question their legitimacy as a selection test.

Overall, the results of this study illustrate a fundamental problem faced by practitioners in the use of concurrent data collection strategies with personality tests. Cutoffs established based upon incumbent samples may vastly underestimate the pass rates for applicants, which could have implications for the cost and utility of selection systems using personality tests.

References

- Alliger, G. M., Lilienfeld, S.O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science*, 7, 32–39.
- Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 66, 1–6.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26.
- Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology*, 69, 25–39.
- Cascio, W. F. (1998). *Applied psychology in human resource management* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional guidelines. *Personnel Psychology*, 41, 1–24.
- Cascio, W. F., Outtz, J. L., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, 4, 233–254.
- Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology*, 47, 847–860.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cook C., Heath, F., Thompson, R. L., & Thompson, B. (2001). Score reliability in web- or internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, 61, 697–706.

- Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, 6, 343–359.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Chicago: Holt, Rinhart, and Winston, Inc.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349–354.
- Cunningham, M. R., Wong, D. T., & Barbee, A. P. (1994). Self-presentation dynamics on overt integrity tests: Experimental studies of the Reid Report. *Journal of Applied Psychology*, 79, 643–658.
- Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, 16, 81–106.
- Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). The validity of non-cognitive measures decays when applicants fake. *Academy of Management Proceedings*, Cincinnati, Ohio.
- Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, 16, 1–23.
- Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology*, 84, 155–166.
- Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe–Crowne social desirability scale. *Educational and Psychological Measurement*, 53, 417–424.
- Guion, R. M., & Cranny, C. J. (1982). A note on the concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 67, 239–244.
- Hatrup, K., O'Connell, M. S., & Wingate, P. H. (1998). Prediction of multidimensional criteria: Distinguishing task and contextual performance. *Human Performance*, 11, 305–320.
- Hatrup, K., O'Connell, M. S., & Labrador, J. R. (2005). Incremental validity of locus of control after controlling for cognitive ability and conscientiousness. *Journal of Business and Psychology*, 19, 461–481.
- Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, 91(1), 9–24.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74, 167–184.
- Hoffman, C. C., & Thornton, G. C. (1997). Examining selection utility when competing predictors differ in adverse impact. *Personnel Psychology*, 50(2), 455–460.
- Hough, L. M., Eaton, N. K., Dunnette, M. D., & Kamp, J. D. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, 75, 581–595.
- Johnson, C. E., Wood, R., & Blinks, S. F. (1988). Spuriousness and spuriousness: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61, 153–162.
- Kehoe, J. F., & Olson, A. (2005). Cut scores and employment discrimination litigation. In F. Landy (Ed.) *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 410–449). San Francisco, CA: Jossey-Bass.
- McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, 85(5), 812–821.
- Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57–81). Mahwah, NJ: Lawrence Erlbaum Associates.
- Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. III (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, 88, 348–355.
- O'Connell, M.S. (2000). *Meta-analysis of the Select Assessment™ for Manufacturing system*. Technical Report: Select International, Inc.
- O'Connell, M. S., Doverspike, D., Gillikin, S., & Meloun, J. M. (2001). Computer Anxiety: Effects on Computerized Testing and Implications for e.Cruiting. *Journal of e.Commerce and Psychology*, 1, 25–39.
- O'Connell, M.S., & Kato, M. (2001). *Reliability analysis for all scales used in the Select Assessment™ for Manufacturing system based on a normative sample*. Technical Report, Select International, Inc.
- O'Connell, M. S., McDaniel, M. A., Grubb, W. L., Hartman, N. S., & Lawrence, A. (2002). *Incremental validity of situational judgment tests for task and contextual job performance*. Symposium presented at the 17th annual meeting of the Society for Industrial and Organizational Psychology, Toronto, Canada.
- O'Connell, M. S., & Smith (1999). *Normative and validity results for hourly worker assessment scales*. Technical report, Select International, Inc.
- Ones, D. S., & Viswesvaran, C. (2003). Job-specific applicant pools and national norms for personality scales: Implications for range-restriction corrections in validation research. *Journal of Applied Psychology*, 88(3), 570–577.
- Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). The role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81, 660–679.
- Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, 78, 579–703.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598–609.
- Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology*, 56, 733–752.
- Pulakos, E. D., Schmitt, N., Dorsey, D. W., Arad, S., Hedge, J. W., & Borman, W. C. (2002). Predicting adaptive performance: Further tests of a model of adaptability. *Human Performance*, 15(4), 299–324.
- Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, 83, 634–644.
- Saldago, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, 82(1), 30–43.
- Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, 78(6), 966–974.
- Schneider, B. (1987). The people make the place. *Personnel Psychology*, 40(3), 437–453.
- SIOP. (1987). *Principles for validation and use of personnel selection procedures* (3rd ed.), College Park, MD: Society for Industrial and Organizational Psychology.
- Smith, D. B., Hanges, P. J., & Dickson, M. W. (2001). Personnel selection and the five-factor model: Reexamining the effects of

- applicant's frame of reference. *Journal of Applied Psychology*, 86(2), 304–315.
- Snell, A. F., Sydel, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, 9(2), 219–242.
- Stark, S., Chernyshenko, O. S., Chan, K., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, 86, 943–953.
- Weekly, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance*, 17(4), 433–461.
- Stewart, G. L., & Carson, K. P. (1995). Personality dimensions and domains of service performance: A field investigation. *Journal of Business and Psychology*, 9(4), 365–378.
- Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe–Crowne social desirability scale. *Journal of Clinical Psychology*, 28, 191–193.
- Svyantek, D. J., & Ekeberg, S. E. (2001). Defining practical significance through academic and practitioner collaboration. *Current Topics in Management*, 6, 25–50.
- Van Iddekinge, C. H., Raymark, P. H., Eidson, Jr., C. E., & Putka, D. J. (2001, April). *Applicant-Incumbent Differences on Personality, Integrity, and Customer Service Measures*. Poster presented at the 16th Annual Meeting of the Society for Industrial and Organizational Psychology: San Diego, CA.
- Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59(2), 197–210.
- Werner, S. H., Jones, J. W., & Steffy, B. D. (1989). The relationship between intelligence, honesty, and theft admissions. *Educational and Psychological Measurement*, 49, 921–927.